

*Amendments*

*In the Specification:*

Please amend the paragraph between the heading “Brief Description of the Drawings” and the heading “Detailed Description of the Invention” as follows. Amendments are shown with additions underlined and deletions in ~~striketrough~~.

FIG. 1 is a control flow diagram according to flowchart ~~illustrating~~ one embodiment of the invention.

As indicated below, please delete the paragraphs added between the paragraph ending on page 9, line 3 and the subheading beginning on page 9, line 4, which were added via the Amendment filed on December 29, 2003. Amendments are shown with additions underlined and deletions in ~~striketrough~~.

~~As discussed above, the first step of the classifying method is to calculate an Object vector, i.e., an ordered set of a small number of data points or scalars (between 4 and 100, more typically between 5 and 30) that is derived from the data stream, FIG. 1, 110 associated with the Object to be classified. The transformation of the data stream into an Object vector is termed “abstraction,” FIG. 1, 120. The most simple abstraction process is to select a number of points of the data stream. However, in principle the abstraction process can be performed on any function of the data stream. In the embodiments presented below abstraction is performed by selection of a small number of specific intensities from the data stream.~~

~~In one embodiment, the second step of the classifying method is to determine in which data cluster, if any, the vector rests. FIG. 1, 130. Data clusters are mathematical constructs that are the multidimensional equivalents of non-overlapping "hyperspheres" of fixed size in the vector space. The location and associated classification or "status" of each data cluster is determined by the learning algorithm from the training data set. The extent or size of each data cluster and the number of dimensions of the vector space is set as a matter of routine experimentation by the operator prior to the operation of the learning algorithm. If the vector lies within a known data cluster, the Object is given the classification associated with that cluster. FIG. 1, 150. In the most simple embodiments the number of dimensions of the vector space is equal to the number of data points that is selected in the abstraction process. Alternatively, however, each scalar of the Object vector can be calculated using multiple data points of the data stream. If the Object vector rests outside of any known cluster, a classification can be made of atypia, or atypical sample. FIG. 1, 140.~~

After the paragraph that ends on page 18, line 22 and before the ABSTRACT, please insert the following paragraphs.

E. Summary of One Embodiment of the Invention

FIG. 1 is a control flow diagram showing the top level processing of the knowledge discovery engine. Processing begins at step 302 and immediately continues to step 304. In step 304, the KDE 202 processes the chromosome strings 204 using a genetic algorithm. The chromosome strings 204 comprise data strings that are to be analyzed. The genetic algorithm inputs the chromosome strings 204 and for each data string, identifies the chromosome variables contained within the chromosome string 204. The chromosome variables 208 define the variables that the KDE 202 will look for in each chromosome string 204.

The KDE 202 continues to step 306 and creates a lead cluster map, or grouping, for each processed chromosome string by using a pre-defined set of variables. The lead cluster map establishes clusters of data records around centroids in high order dimensional space. The membership of a record to a cluster is determined by Euclidean distance. If the Euclidean distance between a centroid and the record places the record inside a decision hyper-radius, the record belongs to the cluster surrounding the centroid. If the Euclidean distance between the record and any existing centroid is greater than the decision hyper-radius, the record establishes a new centroid and a new cluster. All data regarding the lead cluster mapping of the processed chromosome strings is recorded in the string/cluster database 310.

The KDE 202 continues to step 308 wherein for each lead cluster map, it computes a variance across all of the clusters contained within that lead cluster map and records the variance in the string/cluster database 310. This step determines how homogeneous a given chromosome

string 204 is to a predefined set of chromosome variables. The means for determining cluster homogeneity is a statistical measure of the variability of records belonging to a cluster with respect to specific behaviors, outcomes, attributes or the like. In the preferred embodiment, variance is used as the measure of homogeneity, but this is for convenience. It would be readily apparent to one of ordinary skill in the relevant art to use any statistical measure.

Upon completion of step 308, the KDE 202 determines a best lead cluster map; that is, it determines which lead cluster map is the "best fit" with the given sets of chromosome variables.

The KDE 202 continues to step 314 to determine whether the best lead cluster map is less than an acceptable minimum. The acceptable minimum may either be input by the user, or pre-defined within the KDE 202.

If step 314 determines that the best lead cluster map is less than the acceptable minimum, then processing proceeds to step 316. In step 316, the KDE 202 records its final mapping in a chromosome map 210 and displays the best lead cluster map along with the matching variables.

Returning to step 314, if the KDE 202 determines that the best lead cluster map is not less than the acceptable minimum, the KDE 202 proceeds to step 312.

In step 312, the KDE 202 re-processes each processed chromosome string using the genetic algorithm. The genetic algorithm inputs the data for each processed chromosome string from the string/cluster database 310 and reanalyzes them according to the last set of information. After completing the re-ranking of the processed chromosome strings, the KDE 202 returns to step 306 to create new lead cluster maps for each processed chromosome string. The processing continues as described above.